交互式语音识别系统研究

李新辉 王向东 钱跃良 林守勋

摘要：为了实现大词汇量连续语音识别技术的实际应用，本文提出了交互式语音识别的概念并着重研究其中的各项关键技术。所谓的交互式语音识别，是指为语音识别系统配置一位操作员，该操作员在语音识别过程中对识别系统进行指导监督并修正识别结果。同时，识别系统对交互信息进行学习，根据操作员的指导信息和修正信息对内部模型进行自适应调整，从而提高系统的识别性能。本文的研究工作是对当前大词汇量连续语音识别技术实际应用的发展和创新，具有重要科学技术意义和产业应用前景。同时，对语音识别在其他方向（如实时字幕生成，图书馆音频资料整理等）的应用具有实际的借鉴作用。

关键词：语音识别 交互式语音识别 语音语句提取 汉语候选生成 交互式声学模型自适应

1 引言

语音是人类最自然、最重要的交流方式[1]。因此，在计算机相关技术中，自动语音识别作为一种自然、高效的人机交互方式，长期受到各国政府和研究者的高度关注。近年来，语音识别技术取得了长足的进展。面向特殊应用的中小词汇量语音识别技术已经比较成熟[2, 3]，产生了诸如手机语音拨号系统、电话查询系统等实际应用系统。然而，由于受到背景噪音、方言口音、口语化的自然语音以及语义理解等因素的限制，大词汇量连续语音识别的研究仍然停留在实验室阶段，面向真实场景的大词汇量连续自动语音识别系统性能远远无法满足实际应用要求。

在已有的语音识别技术相关研究中，虽然尚没有明确提出交互式语音识别的概念，但已有一些在语音识别过程中引入交互的研究工作。早期研究的代表单位是美国 IBM 公司、卡内基-梅隆大学（CMU）、密歇根大学（University of Michigan）等。其研究主要集中于语音识别的错误纠正技术，即在一句话识别后由说话人对识别结果的错误进行纠正。系统可同时提供多通道的交互方式，包括单词重新发音（re-speaking）、单词拼写（spelling）、键盘输入、手写输入、笔形设备点击、拖动输入、从前 N 个候选（N-best）中选择等[4-7]。近期研究的代表性工作是日本国立高等工业科技研究院（AIST, National Institute of Advanced Industrial Science and Technology）的“音声订正”（speech repair）系统[8] 。该系统对每个单词给出多个候选，并提供相应的交互界面，允许用户在语音输入的同时或完成之后通过选择候选修正语音识别结果。该研究主要针对无噪声的朗读语音，可以达到实时应用，修正后正确率达96%以上。但该系统只提供用户选择界面，没有其它交互功能，也没有利用用户修正信息进行模型自适应，在会议场景等真实自然语音的情况下性能将有较大下降。总的来说，交互式语音识别的相关研究较少，而且多数集中在对结果的修正上，缺乏利用多种交互手段，以及利用交互信息进行声学模型自适应的研究。

为了将大词汇量连续语音识别技术推向实际应用，本文提出了交互式语音识别的概念，研究交互式语音识别中的关键技术，并构造了一个完整的系统。本文所谓的交互式语音识别，是指：为语音识别系统配置一位操作员，在语音识别过程中由其与系统进行交互。其交互方式主要分为两类：一是根据先验知识或当前说话人语音的特点对系统进行适当的指导，例如指示说话人切换、主题切换，指出说话人性别、方言口音类型，甚至将部分先验语料输入系统等；二是根据听觉对当前语音识别结果进行人工修正。考虑到效率和交互的友好性，这类

交互主要采用候选选择的方式，即对一句话进行识别后，为其中的每个字提供多个候选。当第一候选不是正确结果时，操作员可以在其它候选中进行选择或输入正确的内容来纠正识别错误。在交互式语音识别中，系统不仅可以通过操作员的快速修正来修正识别错误，而且可以根据操作员的指导信息和交互信息对内部模型进行选择和自适应。这样模型更加接近当前说话人的发音特点和语音内容，系统输出的候选越来越准确，操作员的修正效率也越来越高，从而满足实际的应用需求。

本文提出的交互式语音识别系统的流程如图 1 所示。在识别开始前，操作员向系统输入待识别对象的信息和谈论主题信息，系统根据操作员的指导信息选择最匹配的声学模型和语言模型。在语音识别过程中，语音经语句提取模块处理后送语音识别模块识别并生成识别中间结果。候选生成模块对识别中间结果进行处理后生成候选。操作员通过选择候选或终端输入来修正识别错误。同时，系统利用提取到的语音语句和对应的修正后文本对声学模型和语言模型进行自适应调整。

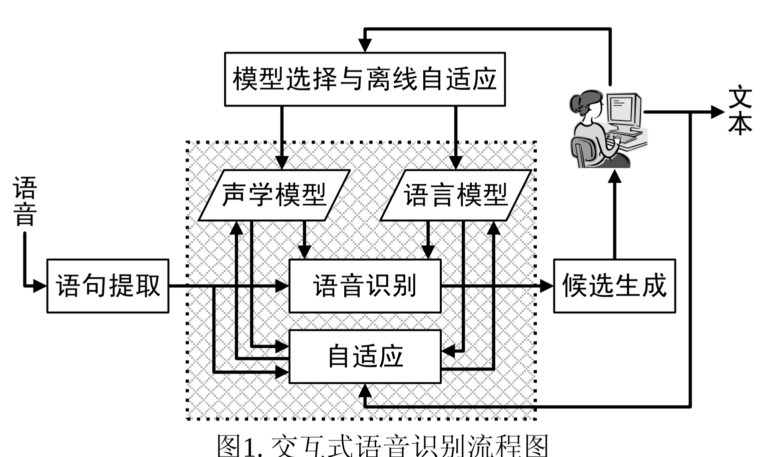


图1. 交互式语音识别流程图

交互式语音识别系统主要包括声学模型、语言模型两个核心模型和语句提取、识别引擎、自适应和候选生成四个核心模块。在本文的研究中，采用了目前世界上较先进的开源 HTK 语音识别引擎[9]，该引擎融合了目前主流的语音识别解码技术。由于生成候选的质量好坏决定了操作员在整个识别过程中的工作效率，同时也决定了交互式语音识别是否能够满足实际应用的需求，本文将研究重点聚焦于如何实时地生成高质量的候选。

2 语音语句提取

在语音识别中，为得到好的结果通常是对一整句话识别完之后输出结果。因此，在对一段语音识别时需要预先提取该段语音中的语句，然后再进行识别。目前主要采用端点检测的方法来提取语音语句。端点检测技术是指从包含语音的一段信号中确定出语音的起始点和结束点。在语音识别中，有效的语音语句提取不仅能减少系统的处理时间、提高系统处理的实时性，而且能排除无声段的噪音干扰，从而使后续的识别性能得以较大提高。

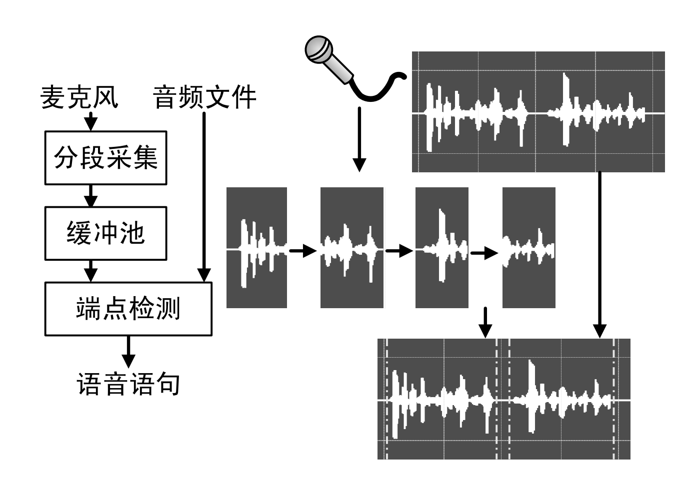


图2. 交互式语音识别中语音语句提取流程图

在交互式语音识别中，识别对象的语音输入既可以是事先录好的音频文件，又可以是实时的语音，语音语句提取模块在这两种情况下都应该能够提取出语音语句。图 2 为交互式语

音识别中的语音语句提取流程图。

在交互式语音识别系统中，对于音频文件输入，系统直接采用端点检测方法对音频文件进行端点检测提取所有的语音语句；对于实时的语音输入，系统实时地采集说话人语音，并对采集到的语音进行端点检测提取语音语句。为了在后一种情况下能够实时地提取语音语句，本文采用分段采集和缓冲池的方法，即每采集一段固定长的音频就把它放到缓冲池中，同时只要缓冲池不为空就从缓存池中拿出一段音频进行端点检测，音频采集与端点检测以同步的方式访问缓冲池。这种方法中，音频固定长度的选取是关键问题：长度过长使得端点检测等待时间过长而影响实时性，长度过短会产生许多无用检测，从而降低系统资源的利用率。本文设定的长度值为 3 秒，因为根据实验统计，大多数情况下，一句话都在 3 秒钟内。

3 汉语候选生成

在交互式语音识别中，候选生成方法直接决定了所生成的候选的质量，而候选的质量好坏决定了操作员在整个识别过程的工作量和工作效率。在国外，主要采用混淆网络生成候选的方法，即利用混淆网络算法（confusion network）[10-12]将词网格压缩成混淆网络来得到候选。使用该方法生成候选必须满足词网格中每条弧对应的对象为一个单独的不可再分割的词。在英语词网格中每条弧对应的词为一个单独的英语单词，因此利用该方法可以生成合适的英语候选。然而，在汉语词网格中每条弧对应的词由一个或多个汉语字组成，每个词可能拆分为两个以上的字（如“中国”，可拆分为“中”和“国”），因此不能利用该方法来生成合适的汉语候选。

通过分析交互式语音识别系统中的需求，我们认为交互式语音识别中的汉语候选生成应满足以下三个约束条件：

（1）具有竞争关系的候选应该属于同一候选列中。这使得操作员只需要在一个候选列中查找正确的候选。

（2）所有候选列应该按照识别时间的先后顺序排列，从而使用户能够按照识别顺序从前往后遍历一次就能够修正所有识别错误。

（3）在每个候选列中，所有候选应该按照识别过程中的得分从高到低排列。得分越高说明该候选为正确词的可能性越大，操作员自上而下查找候选时越容易看到。

3.1 基于字的汉语候选生成方法

为了生成高质量的汉语候选，按照上述提出的汉语候选生成约束条件，我们提出了一种基于字的汉语候选生成方法[13]。在该方法中，首先使汉语词网格对齐，生成对齐网络，然后在对齐网络的基础上将词按字切分生成候选。图 3 为基于字的汉语候选生成示意图。图 3(a)为汉语词网格对齐生成对齐网络；图 3(b)为对齐网络按字切分生成基于字的候选。在本

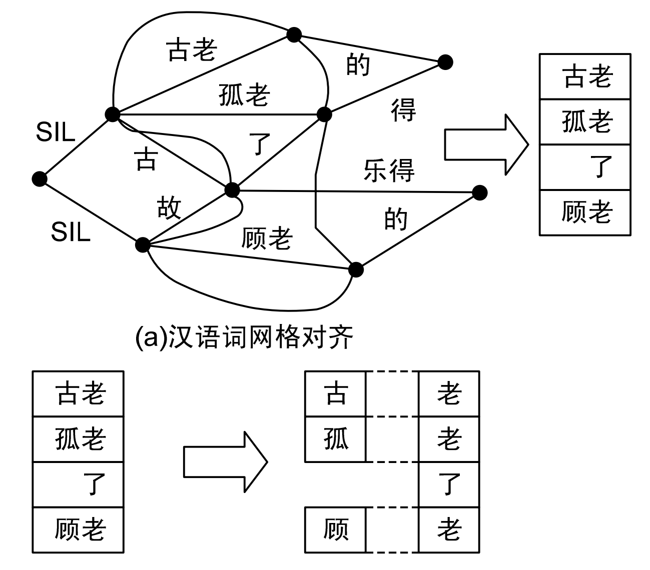


图3. 基于字的汉语候选生成方法示意图

文中，我们分两部分对该方法进行算法描述，一部分为词网格对齐，一部分为字候选生成。在对算法描述之前，我们先给出一些定义：

（1） 汉语词网格

一个汉语词网格由 $L=<N,E>$ 来表示，其中 $N=\{n_0,n_1,n_2,\cdots,n_I\}$ 为汉语词网格中所有结点的集合，$E=\{e_0,e_1,e_2,\cdots,e_J\}$ 为汉语词网格中所有弧的集合。

$\forall n_i\in N$，$t(n_i)$ 表示结点对应的时间。$\forall e_k\in E$，都用一个五元组

$$\{S_{e_k},F_{e_k},W_{e_k},A_{e_k},L_{e_k}\}$$

来表示，其中 S_{e_k} 表示弧 e_k 的起始结点，F_{e_k} 表示弧 e_k 的结束结点，W_{e_k} 表示弧 e_k 上的汉语词，A_{e_k} 表示弧 e_k 的声学概率得分，L_{e_k} 表示弧 e_k 的语言概率得分。

（2） 对齐网络

一个对齐网络由 $E_A=\{E_0',E_1',E_2',\cdots E_K'\}$ 来表示，其中 E_A 为对齐网络中所有对齐类的集合，E_k' 表示第 k 个对齐位置上的弧集合。

（3） 汉字候选

一个汉字候选由 $C=\{C_0',C_1',C_2',\cdots,C_L'\}$ 来表示，C 为候选中所有候选列的集合，$C_l'=\{c_0,c_1,c_2,\cdots,c_K\}$ 表示第 l 个候选列上所有候选集合，$\forall c_k\in C_l'$ 都用一个二元组 $\{W_{c_k},P_{c_k}\}$ 表示，其中 W_{c_k} 表示候选 c_k 对应的候选词，P_{c_k} 表示候选 c_k 对应的得分。

3.1.1 对齐网络的生成

我们可以通过对汉语词网格中的弧进行聚类将汉语词网格对齐，形成对齐网络。聚为一类的弧应满足以下两个条件：（1）每条弧对应词假设的最后一个汉字存在语音相似。（2）弧之间存在时间重叠。

以下为对齐网络生成算法的描述：

步骤 1: 利用前后向算法[10]计算词网格中每条弧 e 的后验概率 $p(e)$。

步骤 2: 将弧集合 E 中的所有的弧，按弧的结束时间 $t(F_{e_k})$ 递增排序，对于结束时间相等的弧，按弧的开始时间 $t(S_{e_k})$ 递增排序。

步骤 3: 初始化 $E_0'=null$，对于 E 中的弧 e，如果 $t(S_{e_k})=0$，则 $E_0'=E_0'\cup e$。

步骤 4: 对于 E 中的每条弧 e_i，$i=0,1,\cdots,J$，假设 $e_{i-1}\in E_I'$：

（a） 若 $t(S_{e_i})=t(S_{e_{i-1}})$ 且 $t(F_{e_i})=t(F_{e_{i-1}})$，则 $E_I'=E_I'\cup e_i$。

（b） 若 $\exists e_i\in E_I'$，使得 $t(S_{e_i})=t(F_{e_i})$，则 $E_{I+1}'=E_{I+1}'\cup e_i$。

$\exists e\in E_I'$，若 $SIM(e,e_i)<SIM(e,e_j)$，则 $E_{I+1}'=E_{I+1}'\cup e$，$E_I'=E_I'\setminus e$。

其中 $SIM(e,e')=sim(c(e),c(e'))\times overlap(e,e')$ 用于计算两条弧之间的竞争程度，$c(e)$ 和 $c(e')$ 分别表示弧 e 和 e' 对应词的最后一个汉字，$sim(\cdots,\cdots)$ 为使用最合适的语音基本公式计算得到的两个汉字的声学相似性，$overlap(e,e')$ 为平滑后的弧 e 和 e' 的时间重叠程度。

（c）　若 $\exists e_j \in E'_K$ 且 $K < I$，使得 $t(S_{e_i}) = t(F_{e_j})$ 且

$$\sum_{l=K+1}^{I} \min_{e' \in E'_l}\{u(e')\} = u(e_i),$$

则 $E'_I = E'_I \cup e_i$。其中 $u(e)$ 表示弧 e 对应汉语词所包含的汉字个数。

（d）　若 $\exists e_j \in E'_K$ 且 $K < I$，使得 $t(S_{e_i}) = t(F_{e_j})$ 且

$$\sum_{l=K+1}^{I} \min_{e' \in E'_l}\{u(e')\} < u(e_i),$$

则 $E'_{I+1} = E'_{I+1} \cup e_i$。

（e）　若 $\exists e_j \in E'_K$ 且 $K < I$，使得 $t(S_{e_i}) = t(F_{e_j})$ 且

$$\sum_{l=K+1}^{I} \min_{e' \in E'_l}\{u(e')\} > u(e_i),$$

则 $E'_H = E'_H \cup e_i$，$K < H \le I$，其中 H 通过以下公式确定：

$$H = \arg\max_{K<H\le I}\{\frac{1}{w(E'_H)} \sum_{e' \in E'_H} SIM(e', e_i)\},$$

其中 $w(E'_H)$ 为 E'_H 中所包含的弧数，$SIM(e, e')$ 与上述定义相同。

步骤 5:　对 E'_k 中每个对齐类，将具有相同汉语词的弧合并成一条弧，其概率值等于合并的弧的后验概率之和。

3.1.2 字候选生成

在对齐网路的基础上，将汉语词切分生成字候选，并对每列候选按照概率得分从高到低排序。

以下为字候选生成算法的描述：

步骤 1:　令 $n = 0$，$m = 0$。

步骤 2:　设 $num = \min_{e' \in E'_n}\{u(e')\}$，

$u(e')$ 同之前定义是一致的，对于 E'_n 中的所有弧 e'_i，$i = 1,2,3\cdots$，都做如下处理：

（a）　若 $u(e') = num$，令候选 c_j 的候选词 $W_{c_j} = Q(We', j)$，候选概率 $P_{c_j} = P(e')$，$C'_{m+j} = C'_{m+j} \cup C_j$，$j = 0,1,\cdots,num-1$，其中 $Q(W_{e'}, j)$ 表示取弧 e' 对应汉语词的第 j 个汉字。

（b）　若 $u(e') > num$，令候选 c_j 的候选词 $W_{c_j} = Q(We', j)$，候选概率 $P_{c_j} = P(e')$，$C'_{m+j+num-u(e')} = C'_{m+j+num-u(e')} \cup C_j$，$j = 0,1,\cdots,u(e')-1$，其中 $Q(W_{e'}, j)$ 表示取弧 e' 对应汉语词的第 j 个汉字。

步骤 3:　$n = n+1$，$m = m + num$，如果 $n < w(E_A)$ 回到步骤 2，否则结束。

步骤 4:　对 C'_k 中对应相同候选词的候选合并为一个候选，其概率值等于合并的候选的概率之和，如果

$$\sum_{c \in C'_k} P_c < 1,$$

令候选 c' 的候选词 $W_{c'} = null$，候选概率

$$P_{c'} = 1 - \sum_{c \in C_k'} P_c \,,$$

$C_k' = C_k' \cup c'$，对合并后的候选按照概率值从大到小排序。

3.2 实验及结果分析

在本实验中，我们对自录的 278 句测试语料进行语音识别，并使用本文介绍的汉语候选生成方法生成候选，最后得出实验结果。实验中用到的声学模型是由 4 万多句的 863 语料和 7 万多句的北方语料训练得到的，语言模型是由 600 多兆的文本语料训练得到的二元语言模型。实验采用的评价标准为：第一候选（1-Best）准确率、前十候选（10-Best）覆盖率、候选平均排名、候选冗余度，其计算公式如下：

1-Best 准确率=1-Best 结果中包含正确字的个数 / 标准答案中字的总个数

10-Best 覆盖率=前 10 个候选中包含正确字的个数 / 标准答案中字的总个数

候选平均排名=正确字在候选中的平均位置

候选冗余度=排在正确字以后的所有字候选之和 / 候选总个数

上述评价标准中，第一候选准确率用来反映语音识别本身的识别性能，即在没有生成候选的情况下，语音识别的正确率；前十候选覆盖率用来反映候选中包含正确词的个数，即能够通过选择候选来修正识别错误的多少；候选平均排名和候选冗余度是站在操作员的角度对候选质量的评价。候选平均排名越靠前，候选冗余度越低，那么操作员查找正确词的速度越快。

表 1 为采用上述评价标准对实验生成的汉语候选进行评价得到的实验结果。

表1　　汉语候选生成实验结果

评价指标	**1-Best 正确率**	**10-Best 正确率**	候选平均排名	候选冗余度
实验结果	76.848%	92.468%	1.65772	77.331%

从表 1 可以看出，使用本文中提出的汉语候选生成方法得到的候选可以修正大部分识别错误。如在本实验中，生成后的汉语候选可以修正多于 15%的识别错误。而且，从候选平均排名来看，在第一个候选和第二个候选中就可以查找到大多数正确的字。

4　交互式声学模型自适应

在交互式语音识别中，生成汉语候选的质量除了受候选生成方法本身的影响外，还受自动语音识别性能的影响。在本文中，利用操作员指导性和修正性的交互信息，提出了基于口音和性别的声学模型选择方法和基于交互信息的有监督声学模型自适应方法。在基于口音和性别的声学模型选择方法中，可根据性别和地域口音事先训练多个声学模型，然后在识别开始前，根据操作员输入的待识别对象信息，为每个说话人选择与之最接近的声学模型。基于交互信息的有监督声学模型自适应方法利用识别过程中已修正的部分识别结果和与之对应的说话人语音，进行有监督声学模型自适应。实验结果表明这两种方法都能够提高自动语音识别的性能，进而提高生成候选的质量。

4.1 基于口音和性别的声学模型选择

为了提高语音识别的性能，进而提高生成的汉语候选的质量，在交互式语音识别中，利用操作员对系统的指导性，本文提出了基于口音和性别的声学模型选择方法。在该方法中我

们根据口音和性别差异训练多个模型，并在识别开始前选择加载与待识别对象发音特点相似的声学模型。在我国不同地域的人对同一个字的发音可能是不同的。如在湖南地区人们习惯将“hu”念成“fu”。此外，男女性别的差异也会造成发音的不同。与男性相比，女性的声音音调通常都较高（即频率高）。因此，本文根据地域口音和性别训练多个声学模型，对每个识别对象根据他的口音和性别选择声学模型，这样能够较大地提高语音识别性能。图 4 为基于口音和性别的声学模型选择流程图。

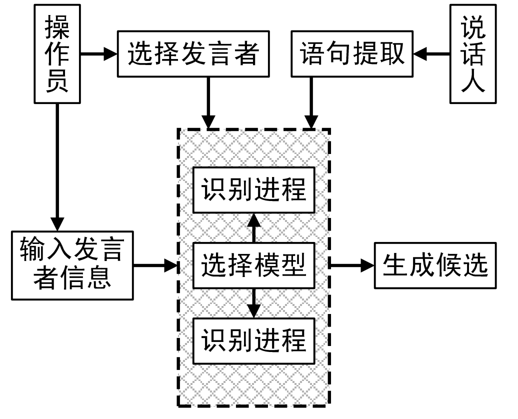


图4. 基于口音和性别的声学模型选择流程图

基于口音和性别的声学模型选择方法主要分为以下几步：

（1）　根据口音和性别训练多个声学模型

根据地域口音和性别对语音语料分类，并对每类语音语料训练一个声学模型。在本文中，我们对本课题研究组积累的北方语音库（带北方口音的普通话）和南方语音库（带南方口音的普通话）按照南北方地域和男女性别进行分类，并对分类后的语音语料分别训练得到四个声学模型（北方男声模型、北方女声模型、南方男声模型、南方女声模型）。

（2）　识别前选择合适声学模型

在识别开始前，操作员输入待识别对象的信息（主要是地域口音、性别），系统根据这些信息为每个待识别对象选择合适的声学模型，并开启相应的识别服务进程。

（3）　识别中实时切换

在识别过程中，当说话人变化时，操作员在系统中标示当前说话人，系统就会将当前说话人的语音语句送到与之对应的识别服务进程进行识别。

4.2 基于交互信息的有监督声学模型自适应

在交互式语音识别中，系统对每句语音识别产生的识别结果都会经过操作员的修正。因此，在交互式语音识别中，利用操作员修正性的交互信息，本文提出了基于交互信息的有监督声学模型自适应方法。在该方法中，我们将已识别的语音和对应的已修正识别结果作为自适应训练语料，对声学模型作有监督自适应。图 5 为基于交互信息有监督声学模型自适应流程图。

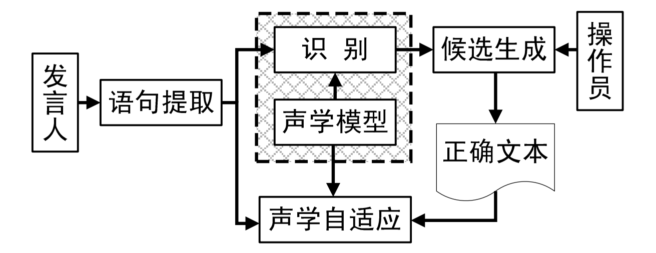


图5. 基于交互信息有监督声学模型自适应流程图

基于交互信息的声学模型自适应方法主要分为以下几步：

（1）　自适应语料收集

在识别过程中，对于每个识别对象我们都为其收集经语音语句提取后的说话语音和经操作员修正后的对应文本信息。

（2）　有监督声学模型自适应

我们利用收集到的语音语料和对应的文本信息，为每个识别对象对应的声学模型作有监督声学模型自适应。该自适应可分为两种：（1）**在线自适应**，当收集到的任何一个识别对象的语音语料超过一定数量（该阈值以句为单位，可设置）时，我们就为其对应的声学模型作有监督自适应；（2）**离线自适应**，在整个识别结束后，我们为每个识别对象对应的声学模型作有监督声学模型自适应，自适应后的声学模型供以后使用。

（3）　声学模型切换

这一步主要针对上文所说的在线自适应。为了使在线自适应后的声学模型能够快速地用于后续的语音识别，提高后续的系统识别性能，我们为在线自适应后的声学模型开启识别服务进程，并在成功开启之后关闭自适应前的声学模型对应的识别服务进程。

4.3 实验及结果分析

4.3.1 声学模型选择

为了验证基于口音和性别的声学模型选择方法对语音识别性能的影响以及对生成候选质量的影响，在本实验中，预先训练了六个声学模型。分别是北方口音男声声学模型、北方口音女声声学模型、北方口音混合声学模型、南方口音男声声学模型、南方口音女声声学模型、南方口音混合声学模型。六个声学模型训练语料的大小都统一为 35750 句，其中混合模型的训练语料中男女声语料各占一半。实验中用到的语言模型是由 600 多兆的文本语料训练得到的二元语言模型。实验测试语料为北方口音 278 句男声语料。实验结果如表 2 所示。

表2　**北方男声测试语料模型选择实验结果**

模型 \ 指标	1-best 正确率	10-best 覆盖率	候选平均排名	候选冗余度
北方男声模型	**77.25%**	**89.35%**	**1.448**	**60.64%**
北方女声模型	57.39%	74.46%	2.135	36.96%
北方混合模型	75.07%	88.40%	1.488	59.98%
南方男声模型	73.09%	86.98%	1.544	57.48%
南方女声模型	49.49%	67.52%	2.754	71.76%
南方混合模型	68.40%	85.70%	1.721	56.40%

上述实验结果中，字体加黑的一栏表示所使用的声学模型实验结果最好。对于北方口音男声测试语料，北方口音男声声学模型测试的实验结果最好。且北方口音声学模型（包括北方男声、北方女声、北方混合声学模型）要好于南方口音声学模型。男声声学模型（包括北方男声、南方男声声学模型）要好于女声声学模型。以上实验结果说明基于口音和性别的声学模型选择能够提高语音识别性能以及候选生成质量。

4.3.2 有监督声学模型自适应

在本实验中，我们将上一实验中的测试语料分成两半。一半用于识别并对识别的结果进行修正，修正后的文本与识别语料一起对北方男声声学模型作自适应。另一半分别用自适应前的北方男声声学模型和自适应后的北方男声声学模型来进行测试，并得到实验结果。整个实验中用到的语言模型是由 600 多兆的文本语料训练得到的二元语言模型。表 3 为北方男声

测试语料的实验结果。

表3 北方男声自适应前后效果对比实验

指标 模型	1-best 正确率	10-best 覆盖率	候选平均排名	候选冗余度
自适应前	77.52%	89.67%	1.523	62.89%
自适应后	84.37%	95.41%	1.428	60.92%

上述实验结果表明，利用修正后的信息对声学模型自适应，采用自适应后的声学模型继续进行识别的结果要好于自适应前的声学模型的识别结果。因此，实验结果表明采用基于交互信息的有监督声学模型自适应能够提高语音识别的性能，以及生成候选的质量。

5 总结和展望

在目前大词汇量连续语音识别无法达到实际应用的情况下，交互式语音识别是对语音识别开启了一种新的应用方式。因此，在交互式语音识别中应充分利用操作员对系统的指导信息和交互信息，提高语音识别的性能，以及候选的质量。交互式语音识别下一步工作有：

（1） 语言模型自适应

语言模型自适应对提高语音识别性能具有较大的作用。在交互式语音识别中，在识别之前可以根据将要谈论的主题搜集与主题相关的语料，然后对语言模型进行事前离线自适应。其次在识别过程中，可根据操作员的修正信息对语言模型进行在线自适应。因此，在将来的工作中利用指导信息和交互信息对语言模型进行自适应具有较好的前景。

（2） 训练更多的区域声学模型

在本文中，我们提到在识别开始前，根据口音和性别事先选择与待识别对象发音相似的声学模型。我国地域广，且各地域的发音不尽相同，几乎所有省份都具有不同口音的普通话。因此，为了提高语音识别性能和提高候选质量，在将来的工作中，可以根据发音不同的地区训练更多声学模型。

总之，交互式语音识别是对目前语音识别一种新的应用方式，可以推广到其他一些应用场景下。

参考文献:

[1] Juang, B.H. and S. Furui, “Automatic Recognition and Understanding of Spoken Language - A First Step toward Natural Human-machine Communication”, Proceedings of IEEE, vol. 88(8): pp. 1142-1165, 2000.

[2] 何湘智, “语音识别的研究与发展”, 计算机与现代化. Vol. 79(3), pp.3-6, 2002.

[3] 姚文冰，姚天任, “稳健语音识别技术研究”, 计算机工程与应用, vol.7, pp.69-71, 2002.

[4] Sharon Oviatt, Phil Cohen, et. al., “Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions”, Human-Computer Interaction, vol.15 (4), pp.263 – 322, 2000.

[5] Suhm, B., Myers, B. , Waibel, A., “Designing Interactive error Recovery Methods for Speech Interfaces”, Proceedings of ACM CHI 1996, Workshop on Designing the User interface for Speech Recognition applications,1996.

[6] Bernhard Suhm, “Empirical Evaluation of Interactive Multimodal Error Correction”, Proc. IEEE Workshop

on Speech recognition and Understanding, pp.583-590, 1997.

[7] Karat, C., Halverson, C., Horn, D., and Karat, “Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems”, Proc. CHI, pp.568-575, 1999.

[8] Jun Ogata, Masataka Goto, “Speech Repair: Quick Error Correction Just by Using Selection Operation for Speech Input Interfaces”, Proc. EuroSpeech, pp.133-136, 2005.

[9] S.Young, J.Jansen, J.Odell, D.Ollason and P.Woodland: The HTK Book, In Entropic Cambridge Research Lab., 1995.

[10] L. Mangu, E. Brill and A. Stolcke, “Finding consensus in speech recognition: word error minization and other application of confusion network,” Computer Speech and Language, vol.14 (4), pp. 373-400, 2000.

[11] L. Mangu, Finding Consensus in Speech Recognition, PhD Thesis, Johns Hopkins University, 2000.

[12] J. Xue and Y.-X. Zhao, “Improved confusion network algorithm and shortest path search from word lattice,” ICASSP 2005, vol.1, pp.853-856, 2005.

[13] Xinhui Li, Xiangdong Wang, Yueliang Qian and Shouxun Lin. Candidate Generation for Interactive Chinese Speech Recognition. Proc. Joint Conferences on Pervasive Computing (JCPC), 2009, 583 - 588

作者简介：

李新辉： 中国科学院计算技术研究所普适计算中心研究生

王向东： 博士，中国科学院计算技术研究所普适计算中心助理研究员 xdwang@ict.ac.cn

钱跃良： 正研级高级工程师，中国科学院计算技术研究所普适计算中心主任

林守勋： 博士，中国科学院计算技术研究所普适计算中心研究员